Mind the Gap: Between-group differences and fair test use

Eyal Gamliel* and Sorel Cahan**

*Behavioral Sciences Department, Ruppin Academic Center, Emek Hefer 40250, Israel. eyalg@ruppin.ac.il
**School of Education, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel

This paper integrates recent meta-analytical findings regarding group differences in job- and educational-related criteria and cognitive ability measures used as predictors in personnel selection and selection to higher education institutions. The findings suggest that cognitive ability measures reveal much higher group differences than the corresponding between-group differences in job- and educational-related criteria. One possible explanation for these differential gaps is that cognitive ability measures are objective and standardized while the typical measures used as job- and-educational related criteria are non-standardized subjective evaluations of job performance and academic achievement. While these findings are consistent with unbiased prediction or over-prediction for lower scoring groups, they imply that selection is biased against them. Implications and future research are discussed.

1. Introduction

The issues of bias in selection and bias in prediction in the context of personnel selection and selection to higher education institutions have recently re-gained momentum. Several articles addressing these issues were published in recent years, most of them presenting theoretical and empirical findings regarding three perspectives: (1) the possible adverse impact caused by the use of various measures (e.g., cognitive aptitude tests and work sample tests), exhibiting substantial differences between majority and minority groups (Bobko, Roth, & Buster, 2005; Callinan & Robertson, 2000; Chung-Yan & Cronshaw, 2002; De Corte & Lievens, 2003; Hough, Oswald, & Ployhart, 2001; Klingner & Schuler, 2004; Roth, Bevier, Bobko, Switzer III, & Tyler, 2001; Roth & Bobko, 2000; Roth, Bobko, & Huffcutt, 2003; te Nijenhuis & van der Flier, 2004); (2) the possible existence of prediction bias regarding majority and minority groups when using various measures, such as cognitive aptitude tests and work samples, for predicting criteria of job performance and success in higher education studies (Cole & Zieky, 2001; Dobson, Krapljan-Barr, & Vielba, 1999; Hough *et al*., 2001; Rotundo & Sackett, 1999; te Nijenhuis, Tolboom, Resing, & Bleichrodt, 2004; te Nijenhuis & van der Flier, 2000); and (3) the possible existence of selection bias regarding majority and minority groups – that is, between-group differences in false-rejection rates (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999; Chung-Yan & Cronshaw, 2002; Cole & Zieky, 2001; De Corte & Lievens, 2003; Roth *et al*., 2003).

The goal of this paper is to integrate the three perspectives by examining the relation between prediction bias and selection bias, and relating it to the recently published meta-analytical data regarding between-group differences in criteria and cognitive ability measures used as predictors. The analysis leads to the following two major conclusions: (1) the conditions for unbiased prediction and unbiased selection are in fact contradictory; (2) personnel selection and selection to educational institutions are probably biased against lower-scoring groups in the United States and in other countries that use similar criteria and cognitive ability measures as predictors.

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd,
9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main St., Malden, MA, 02148, USA

2. Historical background

Ethnic group differences, when using standardized measures of cognitive ability, have been investigated by some of the earliest social science researchers, such as Galton and Thorndike, and this topic continues to receive a great deal of attention (Roth *et al*., 2001). This interest seems warranted given the individual, group, organizational, and social consequences of using measures of cognitive ability in personnel selection and selection to educational institutions.

The differences that exist between majority (e.g., Whites) and minority (e.g., Blacks) groups regarding measures related to cognitive ability have been the topic of many studies. One of the consequences of these differences has been 'adverse impact': lower scoring groups are under-represented in prestigious positions and organizations as well as in higher education programs and institutions. Several researchers tried to account for this phenomenon by examining possible prediction bias (Cleary, 1968). However, typically no prediction bias was found and what was discovered was usually in favor of the lower-scoring group rather than against it (Hartigan & Wigdor, 1989; Jensen, 1980). Nonetheless, even after these studies other members of the psychometric community still felt that personnel selection and selection to higher education institutions were unfair toward members of lower scoring groups and attributed unfairness to selection bias.

The first attempts to formulate and prove this bias and the resulting unfairness in selection involved several definitions of unbiased selection, that is, fair selection. Each of the definitions utilized the four possible results of actual selection decisions based on a predictor, as opposed to the correct selection decisions that would have been made had selection been based on the criterion (Figure 1): (1) Correct Acceptance – An applicant who should have been accepted had selection been based on the criterion is indeed accepted when the acceptance decision is based on the predictor (section I in Figure 1); (2) False Rejection – An applicant who should have been accepted had selection been based on the criterion is actually rejected when the acceptance decision is based on the predictor (section II in Figure 1); (3) Correct Rejection – An applicant who should have been rejected had selection been based on the criterion is indeed rejected when the acceptance decision is based on the predictor (section III in Figure 1); (4) False Acceptance – An applicant who should have been rejected had selection been based on the criterion is actually accepted when the acceptance decision is based on the predictor (section IV in Figure 1).

The theoretical definition of selection bias as a different representation of sub-populations among accepted applicants resulting from the substitution of the predictor for the criterion is accepted by many members of the psychometric community (Cole, 1973; Hartigan & Wigdor, 1989; Linn, 1973; Thorndike, 1971). However, there is disagreement as to the operational definition of selection bias in the psychometric literature. Most suggested definitions explicitly or implicitly rely on the definition of selection errors

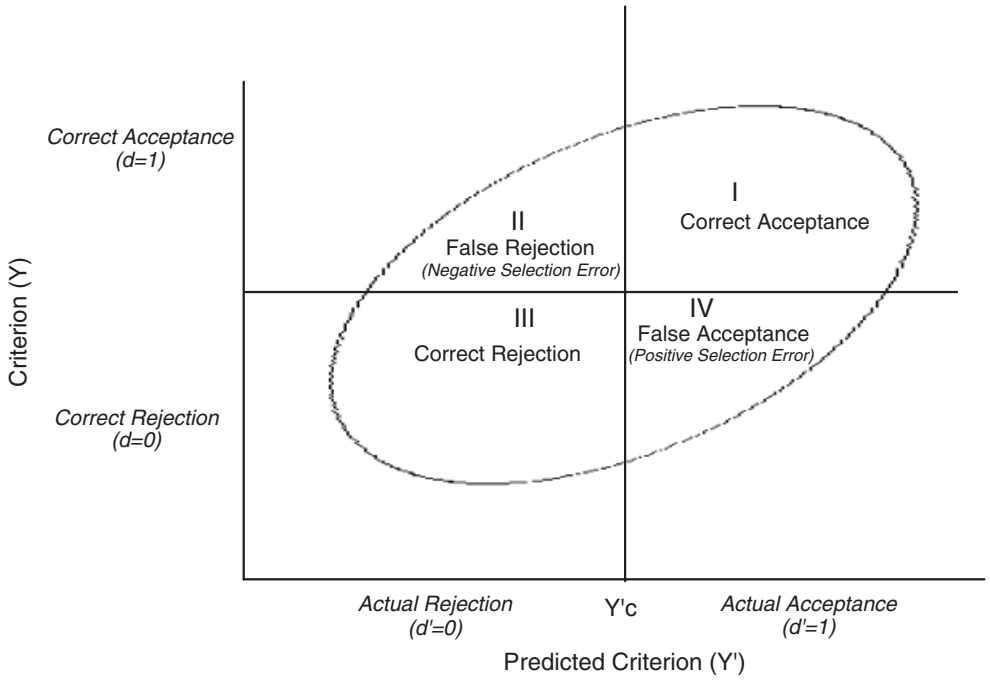

Figure 1. Bivariate distribution of the selection decisions based on criterion *Y* and the selection decisions based on the predictor *X* ($r_{XY}<1$).

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

illustrated in Figure 1. Unbiased selection has typically been defined in terms of *between-group* (minority and majority) equality of the ratio between the probabilities of events represented by the regions in Figure 1: I/(I + II) in the Conditional Probability Model (Cole, 1973); I/(I + IV) in the Equal Probability Model (Einhorn & Bass, 1971; Guion, 1966; Linn, 1973); and (I + IV)/(I + II) in the Constant Ratio Model (Thorndike, 1971).

However, other researchers rejected these definitions on the grounds that they are all based on an internally inconsistent definition of bias. Satisfaction of the condition for unbiased selection, when formulated in terms of success or acceptance probabilities, does not necessarily guarantee satisfaction by the converse probabilities of rejection or failure (Hunter & Schmidt, 1976; Petersen & Novick, 1976). This inconsistency is unavoidable due to the (negative) linear relation between 'percent accepted' (P) and 'percent rejected' ($1-P$), which does not preserve ratio relations despite the apparent absolute nature of the percentage scale (Cahan & Gamliel, 2006).

In spite of the inconsistency of the proposed definitions for unbiased selection, several claims have been raised as to the existence of selection bias in the absence of prediction bias (Chung-Yan & Cronshaw, 2002; Hartigan & Wigdor, 1989; Linn, 1973; Sackett & Wilk, 1994; Thorndike, 1971; Wigdor & Hartigan, 1990). Most of these claims were based on the fact that whenever prediction is not perfect (i.e., practically always), the between-group difference in the predictor is substantially higher than the corresponding difference in the criterion (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999; Chung-Yan & Cronshaw, 2002; Maxwell & Arvey, 1993; Thorndike, 1971).

However, in the absence of a valid definition of selection bias, these claims were not established (Gottfredson, 1994; Hunter & Schmidt, 1976; Petersen & Novick, 1976). The lack of a consistent and valid definition of selection bias impaired, therefore, the investigation of selection bias in social groups and the investigation of the relation between selection bias and prediction bias (Cole & Zieky, 2001; Linn, 1990).

The issue of selection bias is complex and presents many challenges in finding a consistent and valid general definition. Indeed, both the Standards for Educational and Psychological Testing (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999; the Standards) and the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology, 2003; the Principles) provide definitions of predictive bias, but they lack a valid and consistent definition of selection bias.

3. Bias in fixed-*n* selection

While no general valid definition exists, one has been suggested in the specific case of fixed-*n* selection in which the number of qualified applicants (N) exceeds the number of available places (n) (Cahan & Gamliel, 2001). In such cases ($N-n$) applicants must be rejected (Brown, 1980). For every group S, selection bias (SB_S) is consistently and validly defined by the difference between the proportion of accepted applicants from this group when selection is based on a predictor (such as GATB or SAT scores; P'_S), and the correct proportion (P_S) who would have been accepted had selection been based on the criterion, which is unknown at the time of selection (such as job performance or first-year GPA) (Cahan & Gamliel, 2001). Formally

$$SB_S = P'_S - P_S \quad (1)$$

Limiting the definition to fixed-*n* selection situations is not especially problematic because many claims about biased selection have referred to selective positions and organizations as well as to selective educational programs and institutions where the number of qualified applicants exceeds the number of available places. Furthermore, due to self-selection by applicants who are aware of the severe admission requirements, the applicant population itself is selective.

3.1. The relationship between prediction bias and fixed-n selection

Relying on the above definition of selection bias, the following theoretical analysis specifies the conditions for unbiased prediction and unbiased fixed-*n* selection, compares the two conditions, discusses their implication, and illustrates the relationship between the two conditions in recently published meta-analytical data. The analysis is based on two equally sized groups that differ in their average criterion Y (such as job performance measure or first-year GPA): one group with a higher average (henceforth the majority group) and another with a lower average (henceforth the minority group). We will assume that criterion Y and predictor X are measured on an interval scale, that the distributions of the two scores in the two groups are normal with equal variances, and that predictive validity of the criterion Y using predictor X is positive and imperfect, both in the general applicant population and within each group. These assumptions were designed to simplify the analysis and are accepted in the literature regarding prediction and selection, in general (e.g., Lord & Novick, 1968), and in the field of fairness in selection, in particular (e.g., Maxwell & Arvey, 1993; Novick & Petersen, 1976; Petersen & Novick, 1976; Thorndike, 1971). The analysis will further assume that in the total

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

applicant population, the average of X and Y is 0 and that their standard deviations (*SD*) are 1.

3.2. The condition for unbiased prediction

Prediction will be unbiased for the groups defined by any characteristic (e.g., race) if and only if the within-group regression lines coincide, that is, if they are identical to the common regression line (Cleary, 1968). As shown by Thorndike (1971) and further elaborated by Maxwell and Arvey (1993) and Silva and Jacobs (1993), due to the regressive nature of linear prediction, a necessary and sufficient condition for unbiased prediction is

$$\overline{X}_S = \frac{1}{r_{XY}} \times \overline{Y}_S \qquad (2)$$

where the index S indicates group S, and r_{XY} is the predictive validity of the predictor X in the general applicant population.

Equation 2 shows that prediction will be unbiased in group S if and only if the group's mean X score is more extreme than the group's mean Y score by a factor equal to the inverse of the predictive validity. For example, if the predictive validity is .5, and the mean criterion of the majority group is .25, in order for the prediction to be unbiased, the group mean in the predictor should be .5; in other words, twice as extreme as its criterion mean. In this example, the complementary minority group means will be $-.25$ in the criterion and $-.5$ in the predictor. Hence, prediction bias for group S (PB_S) is defined (Linn, 1983; Linn & Hastings, 1984) as

$$PB_S = \overline{X}_S - \frac{1}{r_{XY}} \times \overline{Y}_S \qquad (3)$$

A simultaneous approach to the two complementary groups shows that the condition for unbiased prediction is:

$$\Delta\overline{X} = \frac{1}{r_{XY}} \times \Delta\overline{Y} \qquad [\Delta\overline{X} = \overline{X}_H - \overline{X}_L; \Delta\overline{Y} = \overline{Y}_H - \overline{Y}_L] \qquad (4)$$

where Δ is the between-group mean difference, the index H stands for the higher scoring majority group (H = High), and the index L indicates the lower scoring minority group (L = Low). In other words, for any given difference between the groups' criterion means $(\Delta\overline{Y})$, the smaller the predictive validity, the larger the between-group differences in the predictor $(\Delta\overline{X})$ required for prediction to be unbiased (Maxwell & Arvey, 1993). Because the predictive validity is practically always lower than 1, unbiased prediction requires that the between-group differences in the predictor will be higher than the respective difference in the criterion (i.e., $\Delta\overline{X} > \Delta\overline{Y}$. For example, if the difference between the groups in the criterion is .5 *SD* and the predictive validity is .5, then, in order for prediction to be unbiased, the difference between the groups in the predictor must be 1 *SD*.

Violation of this condition will necessarily lead to prediction bias. If the difference between the groups in the predictor is higher than the ratio between their difference in the criterion and the predictive validity, then the prediction will be biased against the minority group, and vice versa. That is, whenever the between-group difference in the predictor is smaller than this ratio, there will be over-prediction for the minority group and under-prediction for the majority group. In the above example, if the between-group difference in the criterion is .5 *SD* while the difference between them in the predictor is 1.5 *SD*, then the prediction will be biased against the minority group (under-prediction). However, if the difference between the two groups in the criterion is identical to the difference between them in the predictor (e.g., .5 *SD*), then the prediction will be biased in favor of the minority group (over-prediction) and against the complementary majority group (under-prediction). Note that this latter assertion, although statistically correct, is rather counter-intuitive.

3.3. The condition for unbiased selection

According to Equation 1, fixed-*n* selection will be unbiased with respect to group S if and only if the proportion of accepted applicants from this group on the basis of predictor X equals the proportion that should have been accepted had selection been based on criterion Y. Given the above assumptions regarding the distribution of X and Y in the general applicant population and within groups, unbiased selection requires equality between the group differences in the predictor and the criterion. That is, fixed-*n* selection will be unbiased relative to group S if and only if the group mean X score equals the group mean Y score

$$\overline{X}_H = \overline{Y}_H \text{ and } \overline{X}_L = \overline{Y}_L \qquad (5)$$

A simultaneous approach to the two complementary groups shows that the condition for unbiased fixed-*n* selection is

$$\Delta\overline{X} = \Delta\overline{Y} \qquad [\Delta\overline{X} = \overline{X}_H - \overline{X}_L \; ; \; \Delta\overline{Y} = \overline{Y}_H - \overline{Y}_L] \quad (6)$$

That is, fixed-*n* selection will be unbiased only if the mean difference between the two groups in the predictor is identical to their mean difference in the criterion (e.g., .5 *SD*). Violation of this condition will unavoidably lead to selection bias. If the between-group difference in the predictor is smaller than the difference in the criterion (i.e., $\Delta\overline{X} < \Delta\overline{Y}$), selection will be biased in favor of the minority group and against the majority group, and vice versa (Cahan & Gamliel, 2001). For example, if the difference between the groups in the criterion is .5 *SD* while the difference between them in

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

the predictor is 1 *SD*, fixed-*n* selection will be biased against the minority group and in favor of the majority group. Too many applicants from the majority group will be accepted when fixed-*n* selection is actually based on the predictor, as opposed to the correct number who would have been accepted had selection been based on the criterion ($P'_H > P_H$); the opposite holds true for the minority group from which fewer applicants will be accepted compared with the correct number who should have been accepted ($P'_L < P_L$). On the other hand, if the difference between the groups in the criterion is .5 *SD* while their difference in the predictor is only .25 *SD*, fixed-*n* selection will be biased in favor of the minority group and against the majority group. Too many applicants from the minority group will be accepted while too few applicants of the majority group will be accepted.

3.4. Comparison of the conditions for unbiased prediction and unbiased fixed-n selection

The theoretical analysis presented above indicates that the condition for unbiased fixed-*n* selection (i.e., $\Delta\overline{X} = \Delta\overline{Y}$) contradicts the condition for unbiased prediction (i.e., $\Delta\overline{X} > \Delta\overline{Y}$). Therefore, *unbiased prediction will necessarily lead to biased fixed-n selection against the minority group and in favor of the majority group*. In fact, even slight over-prediction for the minority group will still yield bias against it in fixed-*n* selection. For example, if the difference between the groups is .5 *SD* in the criterion and the predictive validity is .5, then there will be a slight over-prediction for the minority group and a selection bias against this group as long as the difference between the groups in the predictor is larger than .5 *SD* and smaller than 1 *SD*.

3.5. Integrating the theoretical analysis with recent empirical findings

The existence and amount of both selection bias and prediction bias are a function of the between-group differences in the predictor and criterion. Table 1 summarizes several relevant research findings. Most of the studies were meta-analyses, performed either in the context of personnel selection or selection to educational institutions, using measures of cognitive ability as predictors. For the most part, research was performed on US data with Whites as the majority group and Blacks as the minority group, but findings regarding the Netherlands are presented as well, with respect to ethnic groups.

The medians of the values presented in Table 1 correspond to between-group differences of .33 *SD* in the criteria and 1.00 *SD* in the predictors (the respective means are .40 and 1.02). These findings are of no surprise: given the common predictive validity of .30 in personnel selection (Bobko, Roth, & Potosky, 1999) and .4–.5 in higher education, the common finding of no prediction bias (Hartigan & Wigdor, 1989; Jensen, 1980; Tenopyr, 1996) requires that the differences between the groups in the predictor will be two to three times higher than the respective differences in the criterion (Chung-Yan & Cronshaw, 2002; Maxwell & Arvey, 1993).

Whenever the between-group differences are substantially higher in the predictors than the performance criteria on a job or in an educational institution, prediction will be unbiased and fixed-*n* selection will be biased against the lower scoring group. The findings presented in Table 1 imply that this will usually be the case: Not enough members from the minority group are accepted relative to the correct number who should have been accepted from this group had selection been based on the criterion of merit.

In order to illustrate these assertions, top–down selection simulations were performed on the criteria and predictors using the data in the meta-analytical studies presented in Table 1 that included information regarding between-group differences on both the cognitive ability measures and the criteria. The simulations used a selection ratio of .5 (one out of two applicants are accepted) and assumed that the minority applicants constitute 25% of the total applicant population (a typical value for all these meta-analytical studies). The simulation also used the assumptions detailed earlier regarding the groups' normal distributions and equal variability. For each meta-analytical study, the between-group differences reported in Table 1 were used to compute the lower scoring group's (i.e., Blacks) selection and prediction biases, using Equations 1 and 3, respectively (in calculating prediction bias predictive validity was set at .30). Table 2 presents the lower-scoring group's (Blacks) estimated selection bias and prediction bias values for the meta-analytical studies presented in Table 1.

Table 2 clearly indicates that in all the meta-analytical studies, the between-group differential gap results in a selection bias against the lower scoring group (Blacks). A median of 23% of the group's applicants are actually accepted when selection is based on the cognitive ability measure vs a median of 41% who would have been accepted had selection been criteria based. On the other hand, prediction was usually biased in favor of this group (in four out of the five studies). These results further emphasize the importance of the conceptual and empirical distinction between the two uses of test scores: prediction and selection. The use of cognitive ability measures as predictors is expected to result in selection bias against lower scoring groups (e.g., Blacks) whereas prediction is expected to favor them. That is, substituting cognitive tests as predictors for the criteria

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

Table 1. The standardized mean difference in predictors of cognitive ability measures and criteria in several studies (*N* indicates number of participants; *K* indicates the number of studies)

Study	Context	Groups	Difference in predictor/s[a]	Difference in criterion/criteria
Schmitt *et al.* (1997) ($K = 14$)	Personnel	White–Black	1.00	.45
Roth *et al.* (2001)	A. Personnel	White–Black	A. .99	—
A. ($K = 34$; $N = 464{,}201$)	B. Military		B. 1.10	
B. ($K = 22$; $N = 387{,}705$)	C. Education I		C. 1.12	
C. ($K = 48$; $N = 5{,}378{,}539$)	D. Education II		D. 1.00	
D. ($K = 38$; 3,007,284)				
Roth and Bobko (2000) ($N = 7498$)	Education	White–Black	—	.43
Rotundo and Sackett (1999) ($N = 23{,}316$)	Personnel	White–Black	.86–.98	.07–.37
Schmitt *et al.* (1996) ($K = 16$, $N = 7590$)	Personnel	White–Black	.83	.15, .33, .38
Martocchio and Whitener (1992) ($K = 10$; $N = 1535$)	Personnel	White–Black	.38	.16
Chung-Yan and Cronshaw (2002) ($K = 149$ studies for *X* and 202 studies for *Y*)	Personnel	White–Black	1.01	.32
McCornack (1983) ($N = 4463$)	Education	White–Black	1.27	.81
Willingham *et al.* (2002) ($N = 7062$)	Education	White–Black	.83	.33
te Nijenhuis and van der Flier (2000)[b] ($N = 1163$)	Education	Dutch vs (1) Surinamese/the Netherlands Antillean (2) Turks (3) Moroccan	(1) .9 (2) 1.54 (3) 1.86	(1) .05 (2) .84 (3) .50
te Nijenhuis and van der Flier (1997)[b] ($N = 2128$)	Personnel	Dutch vs (1) Surinamese (2) Antillians (3) N. Africans (4) Turks	(1) .70 (2) .75 (3) 1.20 (4) .95	—

Notes: [a]All predictors were cognitive ability measures. In personnel selection a typical measure is the General Aptitude Test Battery (GATB) while in educational selection a typical measure is the Scholastic Aptitude Test (SAT). [b]The data is these studies is from the Netherlands.

that are unknown at the time of selection causes bias in selection against the lower scoring groups and in favor of the higher scoring groups, creating a situation in which the 'richer' groups get richer at the expense of the poorer groups.

3.6. Possible causes for the differential between-group gaps

The consistent differential between-group gaps in the criteria and cognitive ability measures as predictors may have many causes. One such cause is suggested by the meta-analysis conducted by Roth *et al.* (2003). This analysis of White–Black differences in job performance in the context of personnel selection examined several objective and subjective measures. Their major finding was: 'larger *d*s associated with objective measures of job knowledge than with subjective measures of job knowledge' (p. 702). Table 3 summarizes their main findings.

Thus, whenever the predictors used in personnel selection and selection to higher education institutions are standardized tests of cognitive ability (e.g., GATB or SAT) while the performance criteria are subjective measures of job performance or grades, it is expected that prediction will be either unbiased or biased in favor of lower scoring groups and that selection will be biased against them.

4. Conclusions

Applying the merit principle of distributive justice implies the acceptance of applicants who will perform better than others either at work or in an educational institution. Unfortunately, the job- and educational-related criteria are unknown at the time of selection. Substituting the typical subjective job- and educational-related criteria by objective cognitive ability tests results in selection of fewer applicants from the lower scoring groups than their correct number, which was relatively small to begin with.

It should be noted that this disproportional representation of the lower scoring group is distinct from the implications of the term 'adverse impact.' Adverse

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

Table 2. The expected selection bias (SB) and prediction bias (PB) for the lower-scoring group (Blacks) in several meta-analytical studies assuming several assumptions[a]

Study	Proportion of acceptance when selection is based on		$SB = P'_S - P_S$	PB
	Predictors (P'_S)	Criteria (P_S)		
Schmitt *et al.* (1997)	.20	.39	−.19	.13
Rotundo and Sackett (1999)	.23	.43	−.21	−.14
Schmitt *et al.* (1996)	.25	.41	−.16	.09
Martocchio and Whitener (1992)	.39	.45	−.07	.12
Chung-Yan and Cronshaw (2002)	.20	.40	−.20	.04

Notes: [a]The assumptions are: relative size of 25%; selection ratio of .50; predictive validity .30; predictor and criterion normal distributions within each group and equal variances.

Table 3. Between-group (White and Black) difference in objective and subjective job performance measures (*N* indicates number of participants; *K* indicates the number of studies) (adapted from Roth *et al.*, 2003)

	Objective measures – *d*	Subjective measures – *d*
Quality measures	.27 ($K = 8$; $N = 2538$)	.26 ($K = 10$; $N = 1811$)
Quantity measures	.35 ($K = 3$; $N = 774$)	.12 ($K = 5$; $N = 494$)
Job knowledge	.61 ($K = 10$; $N = 2027$)	.19 ($K = 4$; $N = 1231$)
Absenteeism	.26 ($K = 8$; $N = 1413$)	.17 ($K = 4$; $N = 642$)

Notes: The standardized difference was corrected for attenuation, such that the raw differences between the objective and subjective measures are even higher.

impact considers relatively low representation of protected groups (e.g., minorities) using a measure for selection. For example, the EEOC 4/5 rule for adverse impact dictates that the proportion of applicants admitted on the basis of any predictor *X* should, ideally, be equal across groups, within a certain level of tolerance (± 20%). While it is perfectly legitimate in principle, such an equality approach is antithetical to the merit principle. According to the latter, large between-group differences in mean criteria scores should be reflected in correspondingly large between-group differences in the proportion of acceptance. It is under the latter, merit principle, that the disproportional representation of lower scoring groups is considered as bias in selection, which is an inherent statistical result of any imperfect albeit unbiased prediction.

This paper reiterates the inherent contradiction between the historical perspectives on biases in test use: prediction bias and selection bias. This tension is mentioned in the Standards (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1999) that claim that whenever prediction is unbiased:

> . . . a lower-scoring group will usually have a higher proportion of examinees who are rejected on the basis of their test scores even though they would have performed successfully if they had been selected. (p. 79)

As stated by the Standards, whenever predictive validity is imperfect, proportionately more false-negative decisions are expected in all lower scoring groups, regardless of group membership. As pointed out by our analysis, this is unavoidable. Because unbiased prediction necessarily results in biased selection and vice versa, it is impossible to keep both prediction and selection unbiased. This inherent contradiction further emphasizes the responsibility of a test user to explicitly claim which use s/he wishes to keep unbiased, unavoidably resulting in bias in the other use.

Although the psychometric literature does not offer an agreed upon definition of unbiased selection, it does offer many solutions for apparently biased, and hence unfair, selection procedures. These solutions could be adopted in order to reduce or eliminate the selection bias documented in this paper. One such strategy replaces the predictors revealing high between-group gaps by other predictors that show smaller gaps. Such attempts characterize the recent suggestions to use measures of work samples instead of cognitive tests as predictors in personnel selection (e.g., Callinan & Robertson, 2000; Hough *et al.*, 2001; Klingner & Schuler, 2004) or situational tests for higher education selection (Lievens & Coetsier, 2002). Those wishing to implement these suggestions were cautioned by Bobko *et al.* (2005), who claimed that the between-group differences on work sample tests could reach .70 *SD*. However, the data in Table 1 reveal a between-group average difference of $< .40$ *SD* in the criteria and about 1 *SD* in the predictors. Thus, while work samples probably reflect higher between-group gaps than the gaps in the criteria, they nevertheless reflect lower gaps than those existing in typical cognitive tests. These data further strengthen

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

the need for developing new predictors that reflect lower between-group gaps. Such alternative predictors could be school grades in the context of selection to higher education institutions, personality measures, biodata, or structured interviews in the context of personnel selection. Indeed, these alternative predictors exhibit lower between-group differences (Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). However, in the context of personnel selection, the predictive validity reported for cognitive ability tests (Bobko *et al*., 1999) is higher than the ones reported for personality measures (Salgado, 2003) and interviews (Moscoso, 2000). In addition, in the context of educational selection, the predictive validity reported for cognitive ability tests (Tenopyr, 1996) is higher than the ones reported for previous grades (e.g., Elliott & Strenta, 1988). Thus, choosing predictors for selection seems to involve an empirical dilemma: Predictors with high predictive validity entail large between-group differences (e.g., cognitive ability tests) while predictors with lower between-group differences have lower predictive validity (e.g., personality measures, interviews, and previous grades).

An alternative strategy could maintain the existing predictors and replace the strict top–down selection rule by alternative procedures, such as within-group norming (e.g., Hartigan & Wigdor, 1989) or 'banding' (e.g., Bobko & Roth, 2004; Cascio, Outtz, Zedeck, & Goldstein, 1991). Still, such alternative strategies are expected to impair the efficiency of the selection procedures and result in the selection of applicants who, on the average, perform lower on the criteria. Indeed, because of these likely consequences, as well as for other reasons, the literature has heavily criticized both within-group norming (e.g., Gottfredson, 1994) and 'banding' procedures (e.g., Schmidt, 1991).

Thus, before suggesting alternatives to the existing selection situation by either seeking other predictors (and thereby reducing predictive validity) or substituting the strict top–down selection rule by an alternative one (thereby reducing efficiency), we propose a different conceptual framework. Instead of presenting any alternative as a solution to the adverse impact problem, we suggest addressing the issue of selection bias. Adverse impact, defined as large between-group gaps in predictors, is considered by those endorsing the equality principle as unfair, and by those endorsing the merit principle as a reflection of differential merits. In contrast, the above conceptual formulation of selection bias shows that the between-group gaps on cognitive ability tests used as predictors are artificially inflated relative to the corresponding criteria gaps. Thus, while the merit principle can justify between-group gaps in magnitudes corresponding to the criteria (about 1/3 *SD*), it cannot justify the threefold as much difference in cognitive ability tests used as predictors. Once this formulation of selection bias is fully assimilated, we believe that the alternative predictors or selection procedures will be more easily accepted by the professional community as well as by the public, as they will be presented within the acceptance of the merit principle as a correction of bias caused by the fact that the job- and educational-related criteria are unknown at the time selection is made. Obviously, those who endorse the equality principle will find these solutions even more appropriate.

Using several reasonable assumptions regarding the distribution of the predictors and criteria, this paper presented and discussed the inherent contradiction between unbiased prediction and unbiased selection, expected whenever predictive validity is less than perfect. This paper further estimated the selection bias against lower scoring groups and prediction bias favoring them with respect to empirical meta-analytical data presented in several studies. Empirical research is needed in order to validate this conceptual and empirical contradiction between selection bias and prediction bias using selection simulations performed directly on full criteria and predictor data, using the proposed measure for selection bias (Equation 1) and the measure for predictive bias (Equation 3).

Additional theoretical and empirical research is needed to further examine the relations between the two concepts dealt with in this paper, namely prediction bias and selection bias, and relate them to the concept of measurement bias. The latter was defined by the 2003 Principles as an irrelevant variance that changes true between-group differences in both predictors and criteria. Measurement can be biased in favor of or against lower scoring or higher scoring groups, and reduce or enlarge existing predictive bias or selection bias.

Future research can implement the suggested theoretical analysis on predictors other than cognitive tests and examine the expected selection bias and predictive bias. Several meta-analytical data are available on predictors that typically show smaller between-group differences, such as work samples (Bobko *et al*., 2005), personality constructs (Hough *et al*., 2001), biodata, and structured interviews (Schmitt *et al*., 1997).

Additional research is also needed in order to examine the trade-off between unbiased selection procedures and selection efficiency. Such research could compare this trade-off between personnel selection and educational selection as well as between private organizations and governmental or federal employers. While private organizations may wish to maximize efficiency at any fair cost, governmental or federal employers may be more sensitive to issues of bias in selection.

Finally, the intricate relations between selection bias and prediction bias need to be examined theoretically and empirically in the context of *n*-free selection, where

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

the proportion (or number) of accepted applicants using the predictors need not equal the corresponding values using the criteria. Such an examination requires a valid and consistent definition of selection bias in this context. We hope that the analysis presented in this paper contributes to such future research.

Acknowledgement

A previous version of this paper was presented at the Annual Meeting of the American Educational Research Association (AERA), Montréal, April, 2005.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999) *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Bobko, P. and Roth, P.L. (2004) Personnel Selection with Top-Score-Referenced Banding: On the inappropriateness of current procedures. *International Journal of Selection and Assessment*, **12**, 4, 291–298.

Bobko, P., Roth, P.L. and Buster, M.A. (2005) Work Sample Selection Tests and Expected Reduction in Adverse Impact: A cautionary note. *International Journal of Selection and Assessment*, **13**, 1, 1–10.

Bobko, P., Roth, P.L. and Potosky, D. (1999) Derivation and Implications of a Meta-Analytic Matrix Incorporating Cognitive Ability, Alternative Predictors, and Job Performance. *Personnel Psychology*, **52**, 3, 561–589.

Brown, C. (1980) A Note on The Determination of 'Acceptable' Performance in Thorndike's Standard of fair selection. *Journal of Educational Measurement*, **17**, 3, 203–209.

Cahan, S. and Gamliel, E. (2001) Prediction Bias and Selection Bias: An empirical analysis. *Applied Measurement in Education*, **14**, 2, 109–123.

Cahan, S. and Gamliel, E. (2006) Definition and Measurement of Selection Bias: From constant ratio to constant difference. *Journal of Educational Measurement*, **43**, 2, 131–144.

Callinan, M. and Robertson, I.T. (2000) Work Sample Testing. *International Journal of Selection and Assessment*, **8**, 4, 248–260.

Cascio, W., Outtz, J., Zedeck, S. and Goldstein, I. (1991) Statistical Implications of Six Methods of Test Score Use in Personnel Selection. *Human Performance*, **4**, 4, 233–264.

Chung-Yan, G. and Cronshaw, S.F. (2002) A Critical Re-Examination and Analysis of Cognitive Ability Tests Using The Thorndike Model of Fairness. *Journal of Occupational and Organizational Psychology*, **75**, 4, 489–509.

Cleary, T.A. (1968) Test Bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, **5**, 2, 115–124.

Cole, N.S. (1973) Bias in Selection. *Journal of Educational Measurement*, **10**, 4, 237–255.

Cole, N.S. and Zieky, M.J. (2001) The New Faces of Fairness. *Journal of Educational Measurement*, **38**, 4, 369–382.

De Corte, W. and Lievens, F. (2003) A Practical Procedure to Estimate the Quality and the Adverse Impact of Single-Stage Selection Decisions. *International Journal of Selection and Assessment*, **11**, 1, 89–97.

Dobson, P., Krapljan-Barr, P. and Vielba, C. (1999) An Evaluation of the Validity and Fairness of the Graduate Management Admissions Test (GMAT) Used For MBA Selection in a UK Business School. *International Journal of Selection and Assessment*, **7**, 4, 196–202.

Einhorn, H.J. and Bass, A.R. (1971) Methodological Considerations Relevant to Discrimination in Employment Testing. *Psychological Bulletin*, **75**, 4, 261–269.

Elliott, R. and Strenta, A.C. (1988) Effects of Improving the Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly. *Journal of Educational Measurement*, **25**, 4, 333–347.

Gottfredson, L.S. (1994) The Science and Politics of Race-Norming. *American Psychologist*, **49**, 11, 955–963.

Guion, R.M. (1966) Employment Tests and Discriminatory Hiring. *Industrial Relations*, **5**, 1, 20–37.

Hartigan, J.A. and Wigdor, A.K. (1989) *Fairness in Employment Testing: Validity generalization, minority issues, and the general aptitude test battery*. Washington, DC: National Academy Press.

Hough, L.M., Oswald, F.L. and Ployhart, R.E. (2001) Determinants, Detection and Amelioration of Adverse Impact in Personnel Selection Procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, **9**, 1&2, 152–194.

Hunter, J.E. and Schmidt, F.L. (1976) Critical Analysis of the Statistical and Ethical Implications of Various Definitions of Test Bias. *Psychological Bulletin*, **83**, 6, 1053–1071.

Jensen, A.R. (1980) *Bias in Mental Testing*. New York: Free Press.

Klingner, Y. and Schuler, H. (2004) Improving Participants' Evaluations While Maintaining Validity by a Work Sample–Intelligence Test Hybrid. *International Journal of Selection and Assessment*, **12**, 1–2, 120–134.

Lievens, F. and Coetsier, P. (2002) Situational Tests in Student Selection: An examination of predictive validity, adverse impact, and construct validity. *International Journal of Selection and Assessment*, **10**, 4, 245–257.

Linn, R.L. (1973) Fair Test Use in Selection. *Review of Educational Research*, **43**, 2, 139–161.

Linn, R.L. (1983) Pearson Selection Formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*, **20**, 1, 1–15.

Linn, R.L. (1990) Admissions Testing: Recommended uses, validity, differential prediction, and coaching. *Applied Measurement in Education*, **3**, 4, 297–318.

Linn, R.L. and Hastings, C.N. (1984) Group Differentiated Prediction. *Applied Psychological Measurement*, **8**, 2, 165–172.

Lord, F.M. and Novick, M.R. (1968) *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Martocchio, J.J. and Whitener, E.M. (1992) Fairness in Personnel Selection: A meta-analysis and policy implications. *Human Relations*, **45**, 5, 489–506.

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

Maxwell, S.E. and Arvey, R.D. (1993) The Search for Predictors with High Validity and Low Adverse Impact: Compatible or incompatible goals? *Journal of Applied Psychology*, **78**, 3, 433–437.

McCornack, R.L. (1983) Bias in the Validity of Predicted College Grades in Four Ethnic Minority Groups. *Educational and Psychological Measurement*, **43**, 2, 517–522.

Moscoso, S. (2000) Selection Interview: A review of validity evidence, adverse impact and applicant reactions. *International Journal of Selection and Assessment*, **8**, 4, 237–247.

Novick, M.R. and Petersen, N.S. (1976) Towards Equalizing Educational and Employment Opportunity. *Journal of Educational Measurement*, **13**, 1, 77–88.

Petersen, N.S. and Novick, M.R. (1976) An Evaluation of Some Models for Culture-Fair Selection. *Journal of Educational Measurement*, **13**, 1, 3–29.

Roth, P.L., Bevier, C.A., Bobko, P., Switzer III, F.S. and Tyler, P. (2001) Ethnic Group Differences in Cognitive Ability in Employment and Educational Settings: A meta-analysis. *Personnel Psychology*, **54**, 2, 297–330.

Roth, P.L. and Bobko, P. (2000) College Grade Point Average as a Personnel Selection Device: Ethnic group differences and potential adverse impact. *Journal of Applied Psychology*, **85**, 3, 399–406.

Roth, P.L., Bobko, P. and Huffcutt, A.I. (2003) Ethnic Group Differences in Measures of Job Performance: A new meta-analysis. *Journal of Applied Psychology*, **88**, 4, 694–706.

Rotundo, M. and Sackett, P.R. (1999) Effect of Rater Race on Conclusions Regarding Differential Prediction in Cognitive Ability tests. *Journal of Applied Psychology*, **84**, 5, 815–822.

Sackett, P.R. and Wilk, S.L. (1994) Within-Group Norming and Other Forms of Score Adjustment in Preemployment Testing. *The American Psychologist*, **49**, 11, 929–954.

Salgado, J.F. (2003) Predicting Job Performance Using FFM and Non-FFM Personality Measures. *Journal of Occupational and Organizational Psychology*, **76**, 3, 323–346.

Schmidt, F.L. (1991) Why all Banding Procedures in Personnel Selection are Logically Flawed. *Human Performance*, **4**, 4, 265–277.

Schmitt, N., Clause, C.S. and Pulakos, E.D. (1996) Subgroup Differences Associated with Different Measures of Some Common Job-Relevant Constructs. In: Cooper, C.L. and Robertson, I.T. (eds), *International Review of Industrial and Organizational Psychology*, Vol. 11. Chichester, UK: Wiley, pp. 115–139.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L. and Jennings, D. (1997) Adverse Impact and Predictive Efficiency of Various Predictor Combinations. *Journal of Applied Psychology*, **82**, 5, 719–730.

Silva, J.M. and Jacobs, R.R. (1993) Performance as a Function of Increased Minority Hiring. *Journal of Applied Psychology*, **78**, 4, 591–601.

Society for Industrial and Organizational Psychology. (2003) *Principles For The Validation and Use of Personnel Selection Procedures* (4th edn). Ohio: Society for Industrial and Organizational Psychology.

te Nijenhuis, J., Tolboom, E., Resing, W. and Bleichrodt, N. (2004) Does Cultural Background Influence the Intellectual Performance of Children from Immigrant Groups?: The RAKIT intelligence test for immigrant children. *European Journal of Psychological Assessment*, **20**, 1, 10–26.

te Nijenhuis, J. and van der Flier, H. (1997) Comparability of GATB Scores for Immigrants and Majority Group Members: Some Dutch findings. *Journal of Applied Psychology*, **82**, 5, 675–687.

te Nijenhuis, J. and van der Flier, H. (2000) Differential Prediction of Immigrant Versus Majority Group Training Performance Using Cognitive Ability and Personality Measures. *International Journal of Selection and Assessment*, **8**, 2, 54–60.

te Nijenhuis, J. and van der Flier, H. (2004) The Use of Safety Suitability Tests for the Assessment of Immigrant and Majority Group Job Applicants. *International Journal of Selection and Assessment*, **12**, 3, 230–242.

Tenopyr, M.L. (1996) The Complex Interaction Between Measurement and National Employment Policy. *Psychology, Public Policy, and Law*, **2**, 2, 348–362.

Thorndike, R.L. (1971) Concepts of Culture-Fairness. *Journal of Educational Measurement*, **8**, 2, 63–70.

Wigdor, A.K. and Hartigan, J.A. (1990) The Case for Fairness. *Society*, **27**, 3, 12–16.

Willingham, W.W., Pollack, J.M. and Lewis, C. (2002) Grades and Test Scores: Accounting for observed differences. *Journal of Educational Measurement*, **39**, 1, 1–37.

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd